Ulrike Kuhl
Bielefeld University, Bielefeld, Germany

Mark T. Keane
University College Dublin, Dublin, Ireland

More Complex Designs

- We have...looked at simple 2-GP Expt
- Now we consider more complex designs
- With new concerns that arise

Six Steps to Heaven…

Motivation

Design

Materials & Procedure

Piloting

Data Collection & Analysis

Results

Design: Two Groups

Control:
No-Explanation

Experimental:
CF-Explanation

What are we going to show people?

Design: N Groups (N=3)

Design

Control:	Experimental#1: XP-method-1	Experimental#2: XP-method-2

What are we going to show people?

Design: N Groups (N=3)

Design

Control:	Experimental#1: XP-method-1	Experimental#2: XP-method-2

Is this just more people ?

Design: N Groups (N=3)

Design

Control:	Experimental#1: XP-method-1	Experimental#2: XP-method-2
material-set procedure	material-set procedure	material-set procedure

Matching of materials & procedure become harder!

Bird ID App: Classifying Terns

STEP 3

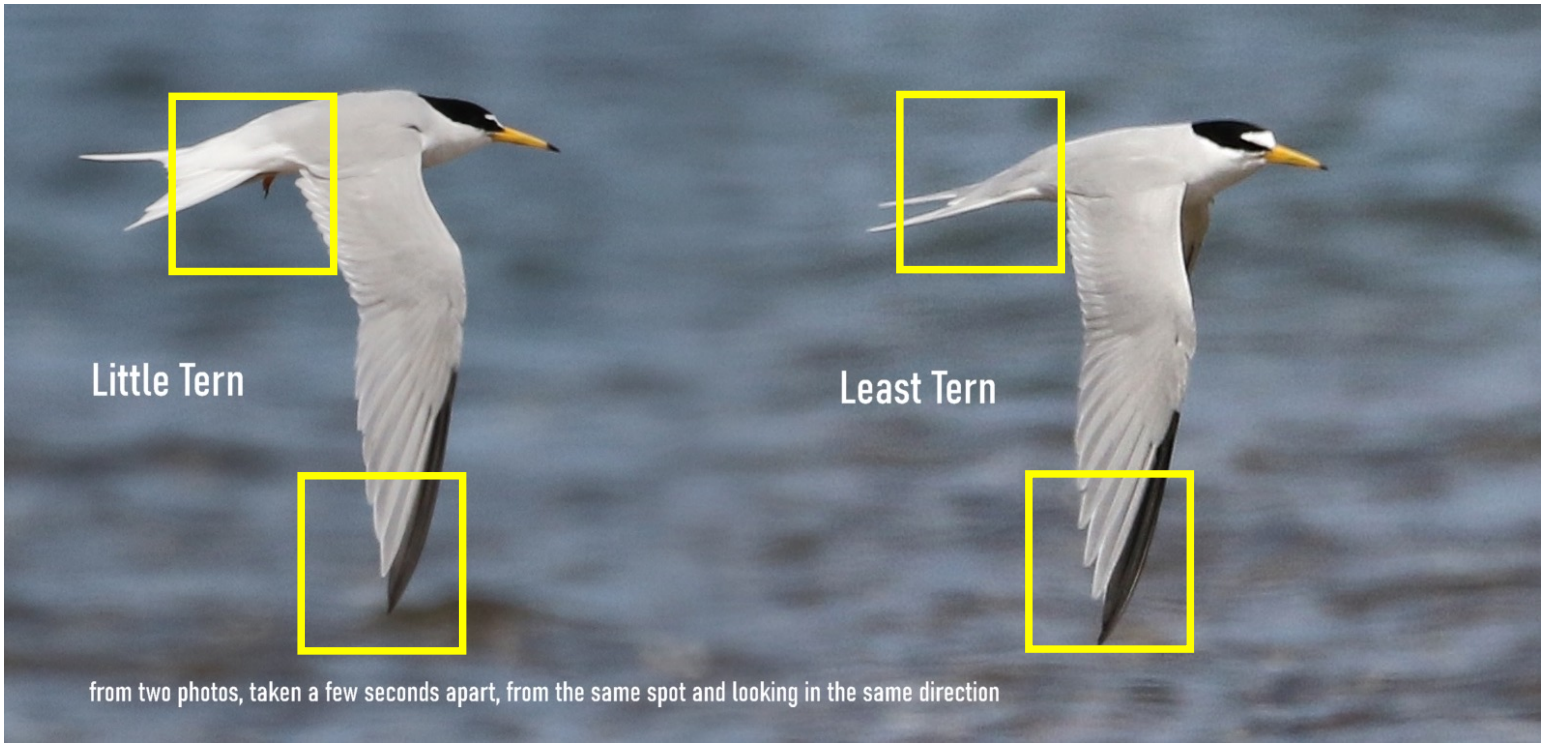
Materials & Procedure

Which is Little, Which is Least?

Bird ID App: Classifying Terns

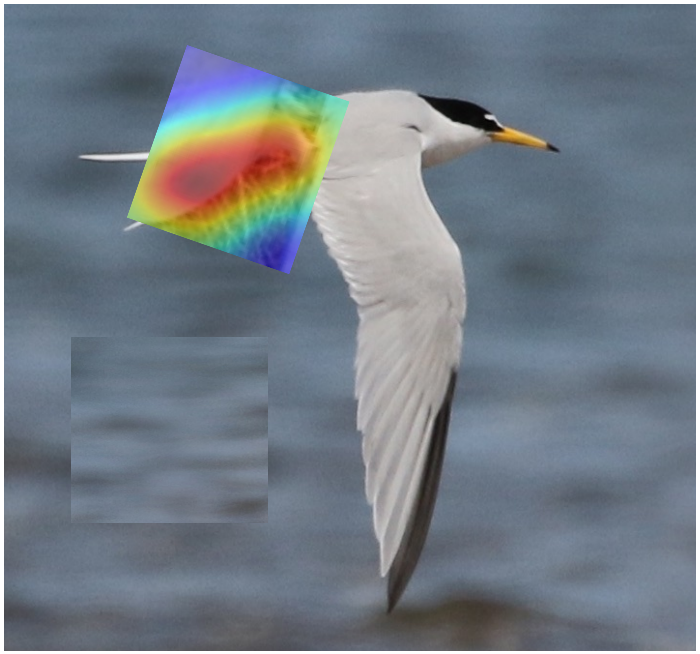
STEP 3

Materials & Procedure



Which is Little, Which is Least?

Method #1: Feature Importance

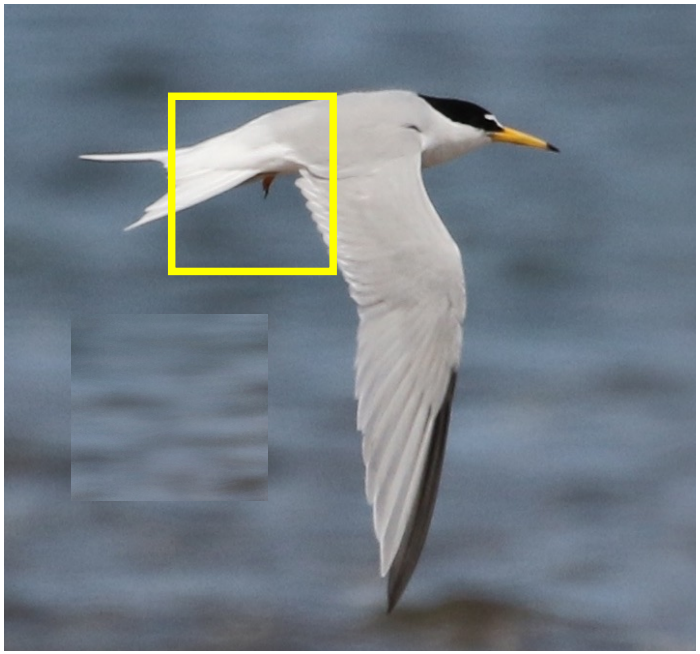


App says this is a:

Little Tern

because of the highlighted area(s) in picture.

Method #2: Counterfactual



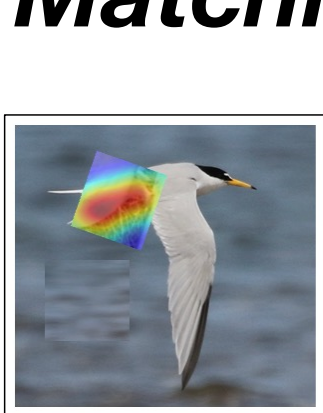
App says this is a:

Little Tern

if the highlighted area was *grey* then it would be a ***Least Tern***.

Matching Issues?

Materials & Procedure

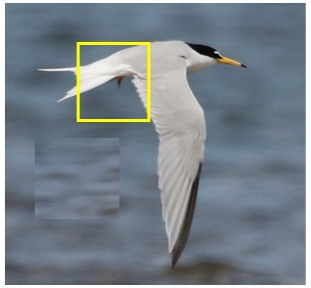


App says this is a:

Little Tern

because of the highlighted area(s) in picture.

V



App says this is a:

Little Tern

if the highlighted area was *grey* then it would be a ***Least Tern***.

If behavior is different, what is it due

~ feature-attribution V counterfactual*

~ the use of salience-box V yellow-box

~ see the white-patch V salience-box

~ confusion over words, “area(s)” V “area”

~ words used in one over the other

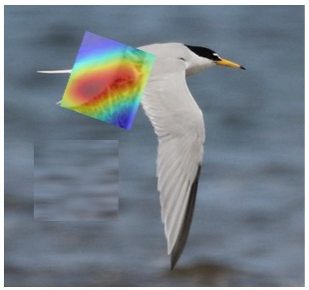
~ poor positioning of the saliency patch

~ color confusions of saliency patch …

Matching Issues?

material-issues
procedure-issue

Materials & Procedure

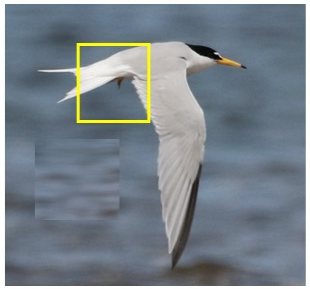


App says this is a:

Little Tern

because of the highlighted area(s) in picture.

V



App says this is a:

Little Tern

if the highlighted area was *grey* then it would be a ***Least Tern***.

If behavior is different, what is it due to?

~ feature-attribution V counterfactual*

~ the use of salience-box V yellow-box

~ see the white-patch V salience-box

~ what people are being asked to do

~ poor positioning of the saliency patch

~ color confusions of saliency patch …

~ use of area(s) versus area

Materials Solutions?

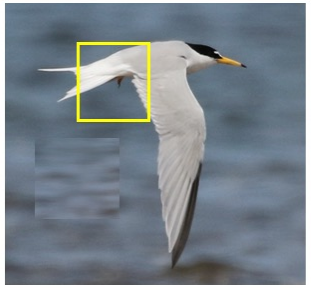
Materials & Procedure

App says this is a:

Little Tern

because of the highlighted area(s) in picture.

V



App says this is a:

Little Tern

if the highlighted area was *grey* then it would be a ***Least Tern***.

Create new variable (h-salience v h-box) and cross that with method; but then you have unbalanced design.

h-salience

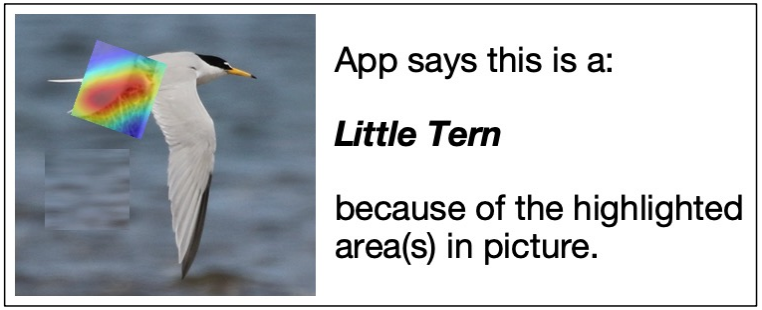
h-box

Oops, 5 conditions!

	h-sal.	h-box	
Control	Expt1.1	Expt1.1	method-1
	Expt2.1	Expt2.1	method-2

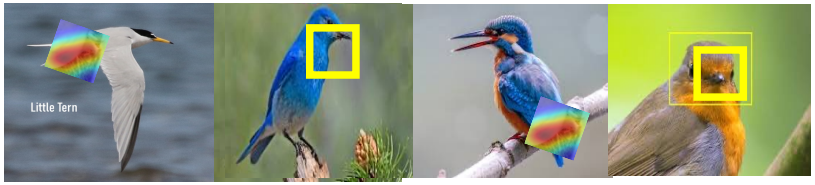
Simple Solution?

Materials & Procedure



V

Take the 4 picture-items and for each person randomly assign salience-boxes to ½ and highlight-boxes to other ½



This will control for this box-variable across methods but will mean we will not know if they have a role to play.

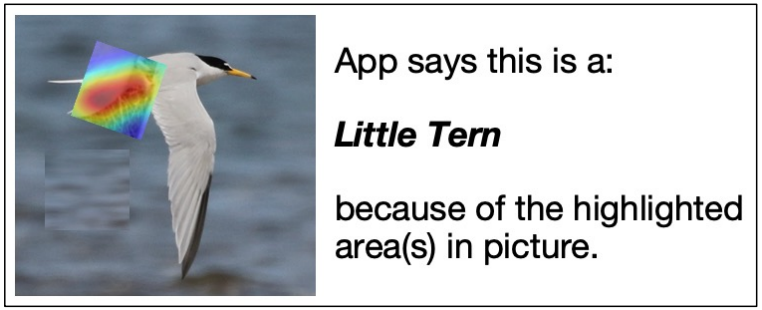
Back to 3 conditions!

balanced
h-box + h-sal

- **Control**
- **Expt1.1** method-1
- **Expt2.1** method-2

Procedural Issues?

Materials & Procedure



V

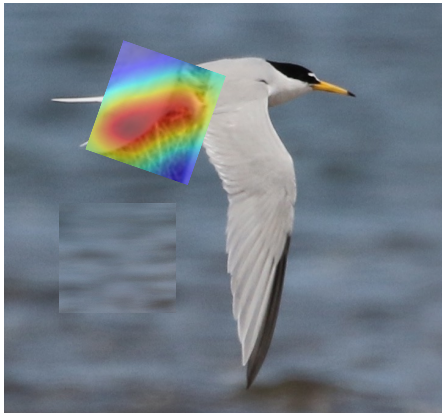
If behavior is different, what is it due to?

~ what people are being asked to do

The counterfactual refers to a region+feature+bird, but the feature one just refers to a region ! Does this matter, is it really just reflecting what the method does?

Procedural Solution?

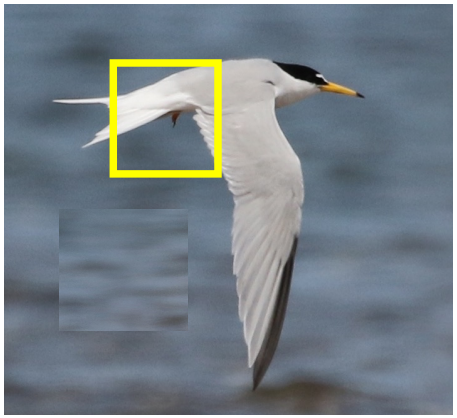
Materials & Procedure



App says this is a:

Little Tern

because the highlighted tail-feature makes it a ***Little Tern***.



App says this is a:

Little Tern

if the highlighted tail-feature was grey it would make it a ***Least Tern***.

Materials & Procedure

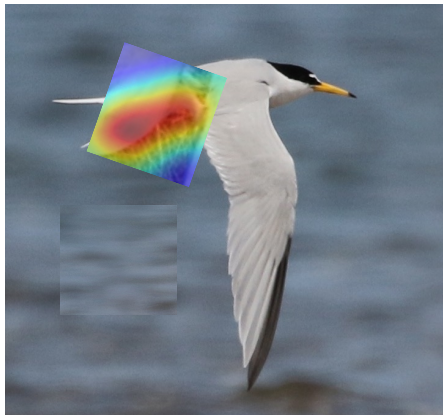
Control

App says this is a:

Little Tern

because the features we see make it a ***Little Tern***.

Experimental#1:
XP-method-1

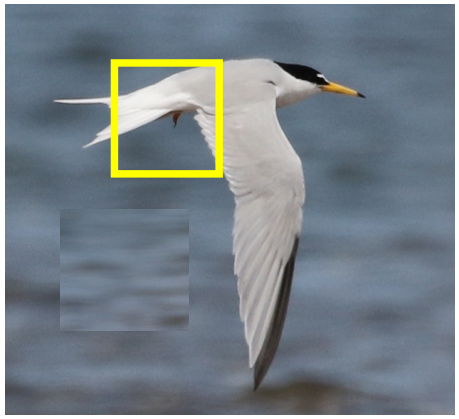


App says this is a:

Little Tern

because the highlighted tail-feature makes it a ***Little Tern***.

Experimental2:
XP-method-2



App says this is a:

Little Tern

if the highlighted tail-feature was grey it would make it a ***Least Tern***.

Procedure: Measures

Materials & Procedure

Piloting

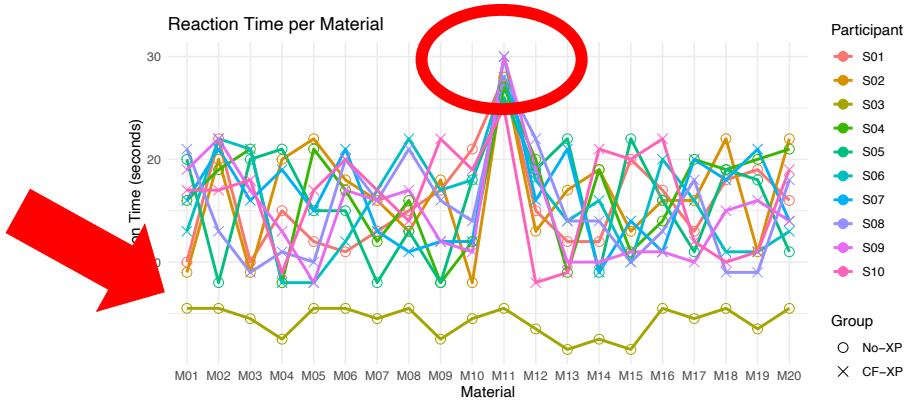
- Image-based learning classification is task (nb. BAC one earlier is also classification)
- We could ask people to check the correctness of the App's classification ***(objective measure)***
- And/Or we could ask if the explanation is helpful/satisfactory etc… ***(subjective measure)***
- Again we could pilot it to test these options.

Data Collection & Analysis

Data Collection & Analysis

After Collecting, Cleaning !

- Legitimate removal of bad data-points/people:
  - speedsters, strange materials (pilot!), attention failures, straight-liners, non-varying responses…



Doing Descriptive Statistics

Data Collection & Analysis

Mean, median, mode, standard deviation…

Make sure to describe your sample!

"We recruited 15 participants, randomly assigned to one of the three groups the ***Control, XP-method-1 and XP-method-2 Groups****. No participants were excluded because reaction times were >3 SDs from the sample mean, there was no straight-lining response patterns (i.e., same response in 5 consecutive trials) and all attention checks were passed.*

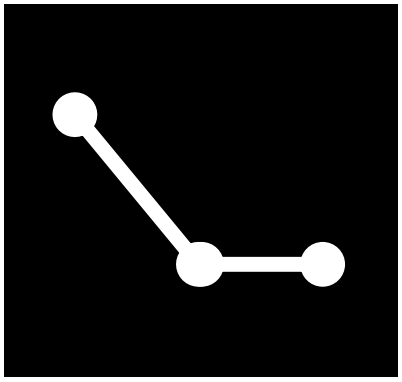
Thus, data from all participants was included in the final analysis for the ***Control Group*** *(n=5, 3 female, median age 35-44y...); the* ***XP-method-1 Group*** *(n=5, 2 female, median age 35-44y...); and the* ***XP-method-2 Group*** *(n=5, female, median age 25-35y...)."*

Starting the Statistics

STEP 5

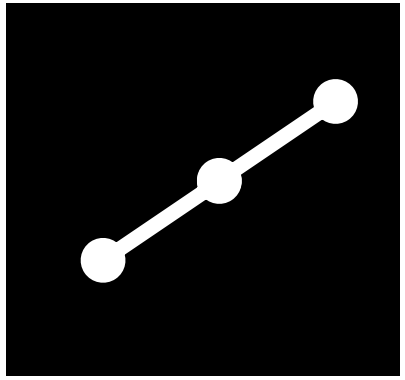
Data Collection & Analysis

Simplest first step is to just look at the means of the groups, this can tell you a lot...



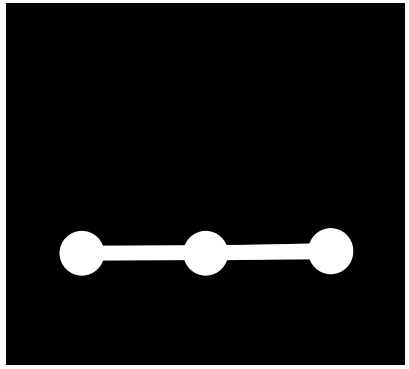
A B C

A > B,C



A B C

trend with C >> A and C > B



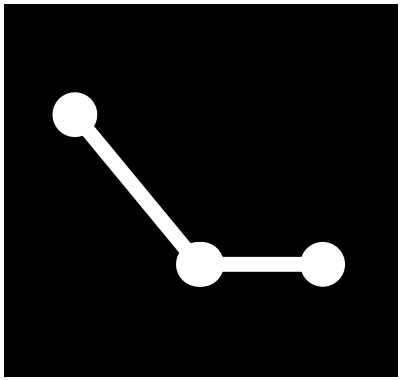
A B C

no difference

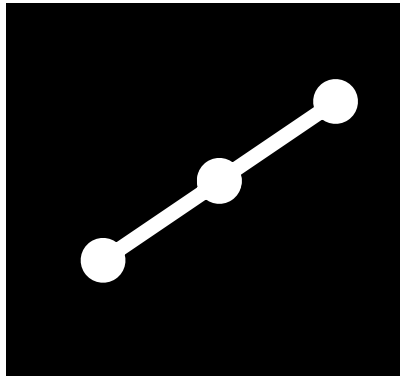
Starting the Statistics

Data Collection & Analysis

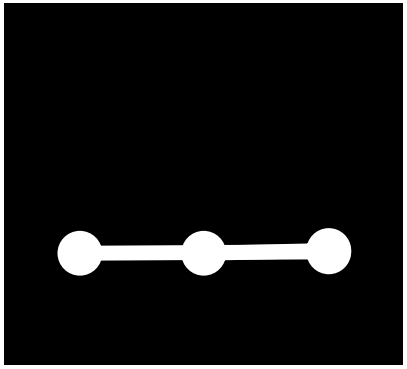
Simplest first step is to just look at the means of the groups, this can tell you a lot...



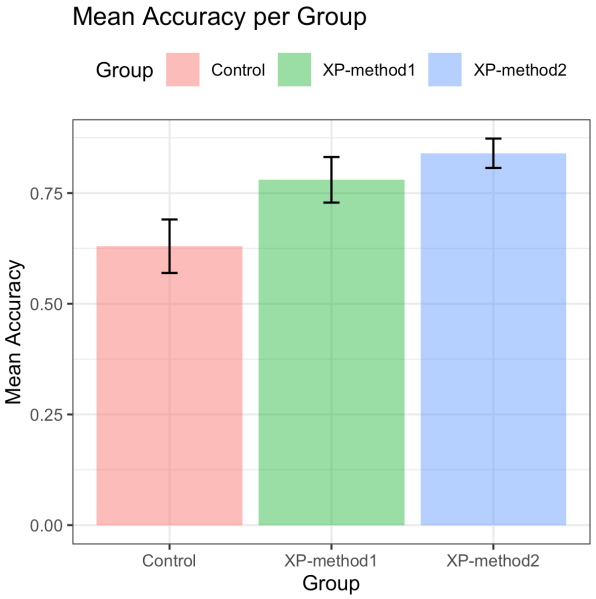
A > B,C



trend with C >> A and C > B



no difference



Accuracy: Looks like XP-m2 > control … at least…?

Data Collection
& Analysis

Design Constrains the Statistical Choices !!

N-group case (that is ND) is a **Kruskal-Wallis Test***.

***Alternative options:**

- One-way ANOVA (note stronger assumptions!)
- Page's L Test for a trend

Design Constrains the Statistical Choices !!

Data Collection & Analysis

N-group case (that is ND) is a **Kruskal-Wallis Test***.

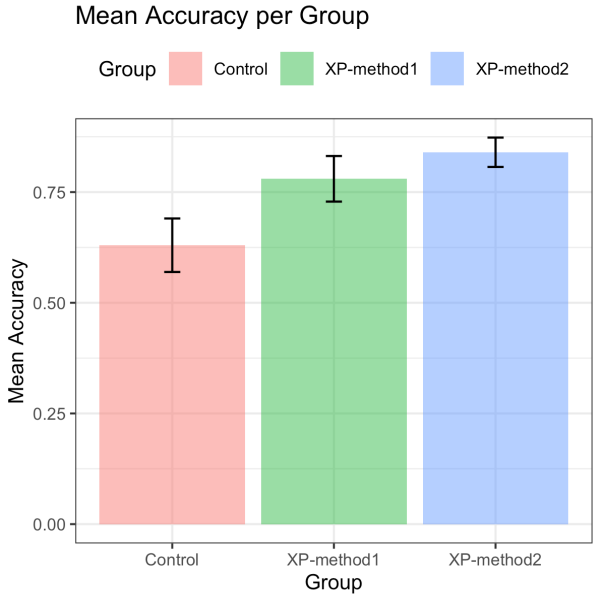
```
        Kruskal-Wallis rank sum test

data:  Accuracy by Group
Kruskal-Wallis chi-squared = 12.438, df = 2, p-value = 0.001991
```

Bingo, something is different!

… but what exactly?

Significant Kruskal-Wallis tells us:

At least one group is different in terms of accuracy – but we cannot say which one, yet!



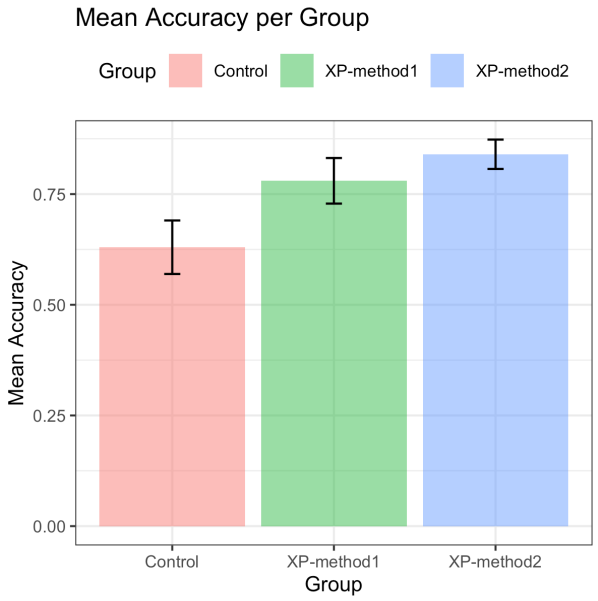
Post-hoc Analysis to the Rescue

Data Collection & Analysis

Follow-up Significant **Kruskal-Wallis** tests with **Dunn's Test** (Dunn, 1964, Technometrics)

???

```
# A tibble: 3 × 9
  .y.      group1     group2       n1    n2 statistic      p   p.adj p.adj.signif
* <chr>    <chr>      <chr>     <int> <int>     <dbl>  <dbl>  <dbl> <chr>
1 Accuracy Control    XP-method1   100   100      2.45 0.0145 0.0434 *
2 Accuracy Control    XP-method2   100   100      3.42 0.000618 0.00185 **
3 Accuracy XP-method1 XP-method2   100   100     0.978 0.328  0.984  ns
```



The Curse of Multiple Comparisons

Data Collection & Analysis

Beware: each statistical test we run carries a chance of error! We might find:

- ***Type I Error*** = false positive: Detect an effect that is not actually there

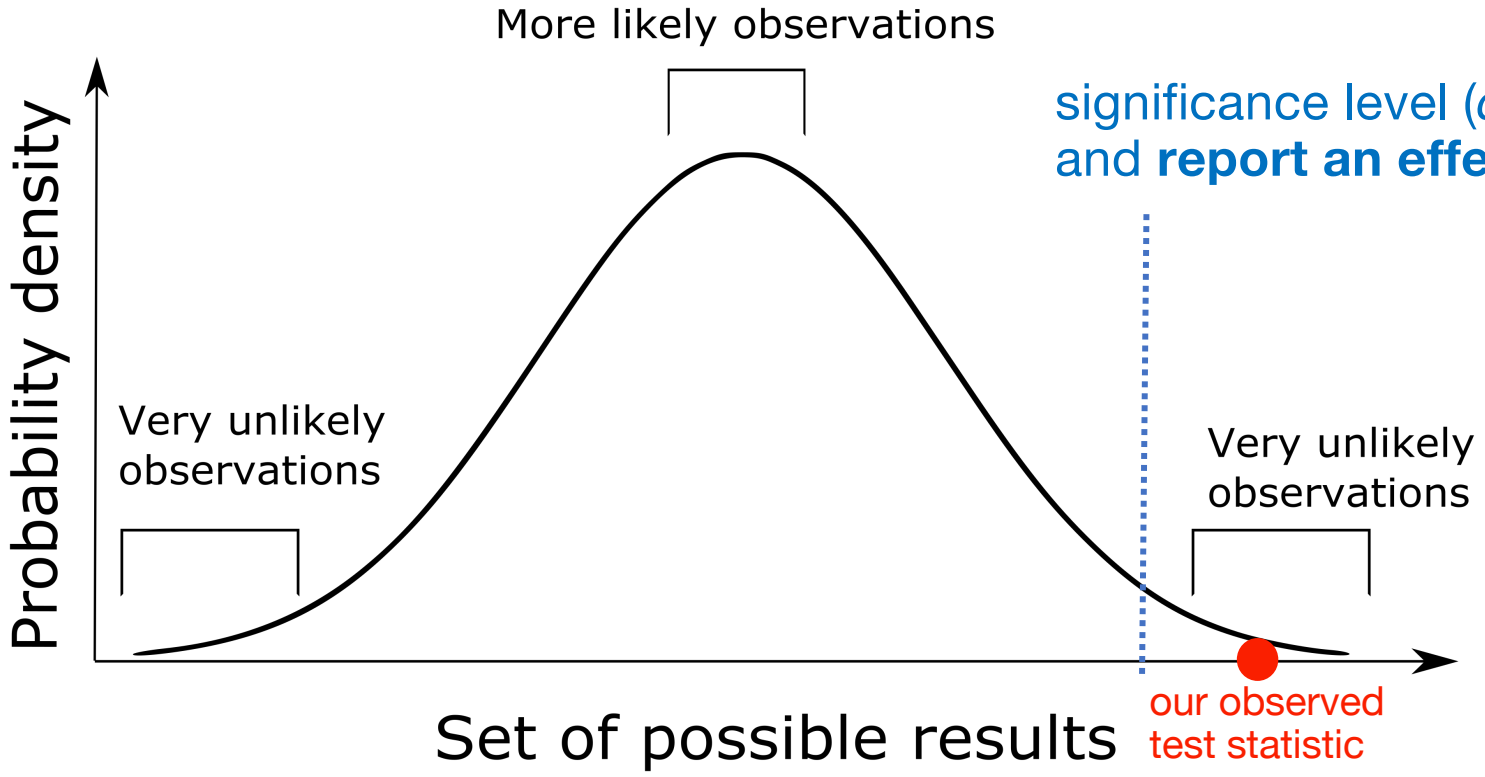
- ***Type II Error*** = false negative: Overlook an effect that is there

The Curse of Multiple Comparisons

Data Collection & Analysis

Beware: each statistical test we run carries a chance of error! We might find:

- ***Type I Error*** = false positive: Detect an effect that is not actually there



significance level (α), and **report an effect if $p < \alpha$**

$\alpha = 0.05$: we accept a 5% probability of getting a false positive

The Curse of Multiple Comparisons

Data Collection & Analysis

An analogy:

drawing marbles, with replacement – what is the chance for getting black?

	Overall chance:
Draw 1: 1/20	5%

https://xkcd.com/882/

The Curse of Multiple Comparisons

Data Collection & Analysis

An analogy:

drawing marbles, with replacement – what is the chance for getting black?

Overall chance:

Draw 1: 1/20

Draw 2: 1/20

9.75%

https://xkcd.com/882/

The Curse of Multiple Comparisons

STEP 5

Data Collection & Analysis

An analogy:

drawing marbles, with replacement – what is the chance for getting black?

Draw 1: 1/20

Draw 2: 1/20

Draw 3: 1/20

Overall chance:

14.26%

https://xkcd.com/882/

The Curse of Multiple Comparisons

Data Collection & Analysis

An analogy:

drawing marbles, with replacement – what is the chance for getting black?

Draw 1: 1/20

Draw 2: 1/20

Draw 3: 1/20

Draw 4: 1/20

Overall chance:

18.55%

https://xkcd.com/882/

The Curse of Multiple Comparisons

Data Collection & Analysis

An analogy:

drawing marbles, with replacement – what is the chance for getting black?

Draw 1: 1/20

Draw 2: 1/20

Draw 3: 1/20

Draw 4: 1/20

.

.

.

Draw 20: 1/20

Overall chance:

64.15%

https://xkcd.com/882/

The Curse of Multiple Comparisons

Data Collection & Analysis

Inflated chance of committing a Type I Error!

Bonferroni-correction

Convention: choose a significance level (α) of 0.05

Carlo Emilio Bonferroni

The Curse of Multiple Comparisons

Data Collection
& Analysis

Inflated chance of committing a Type I Error!

Bonferroni-correction

Convention: ~~choose a significance level (α) of 0.05~~

Adjust significance level:

- divide *α* by the number of comparisons
- Conversely: multiply *p* with number of comparisons

  Example for three comparisons:
  $\alpha_{orig} = 0.05 \dashrightarrow \alpha_{adj} = 0.05 / 3 = 0.0167$

  $p_{orig} = .0145 \dashrightarrow p_{adj} = .0145 * 3 = .0435$

Carlo Emilio
Bonferroni

Post-hoc Analysis to the Rescue

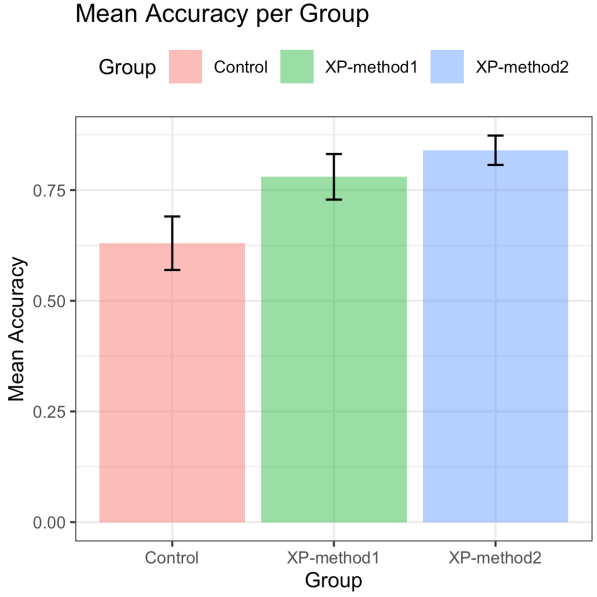
Data Collection & Analysis

Follow-up Significant **Kruskal-Wallis** tests with **Dunn's Test** (Dunn, 1964, Technometrics)

```
# A tibble: 3 × 9
  .y.      group1     group2        n1    n2 statistic        p   p.adj p.adj.signif
* <chr>    <chr>      <chr>      <int> <int>     <dbl>    <dbl>   <dbl> <chr>
1 Accuracy Control    XP-method1   100   100      2.45  0.0145   0.0434  *
2 Accuracy Control    XP-method2   100   100      3.42  0.000618 0.00185 **
3 Accuracy XP-method1 XP-method2   100   100      0.978 0.328    0.984   ns
```

Bingo, something is different!

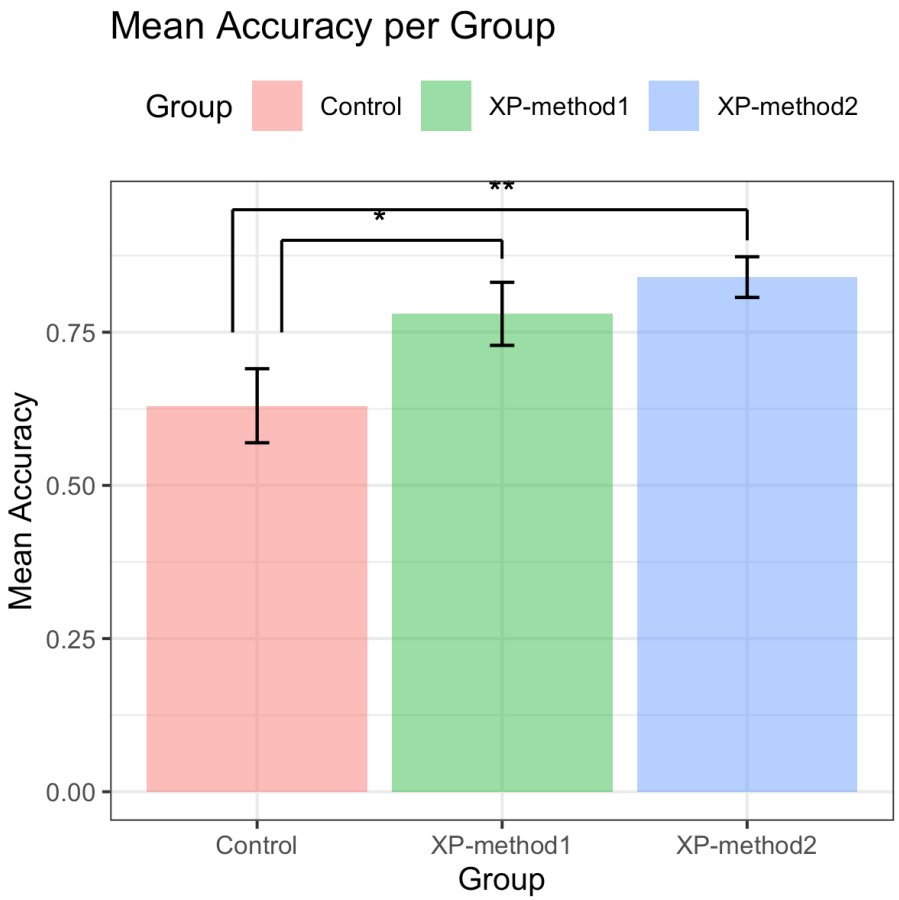
… and now the know!



Put a Nice Bow on It!

STEP 6

Results



A Kruskal-Wallis Test found that the differences in accuracy between the three groups were statistically significant ($H(2)=12.44$, $p=.002$).

Planned pairwise comparisons of the three conditions using Dunn's test indicated that accuracy of control group participants was significantly different from those of the XP-materials-1 Group ($p = .043$) and the XP-materials-1 Group ($p = .002$).
There was no evidence, however, for a difference in accuracy between the two experimental groups ($p = .984$).

Plan for the Day

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END